D0 Run II Farms

M. Diesburg, B.Alcorn, J.Bakken, T.Dawson, D.Fagan, J.Fromm, K.Genser, L.Giacchetti, D.Holmgren, T.Jones, T.Levshina, L.Lueking, L.Loebel-Carpenter, I.Mandrichenko, C.Moore, S.Naymola, A.Moibenko, D.Petravick, M.Przybycien, H.Schellman, K.Shepelak, I.Terekhov, S.Timm, J.Trumbo, S.Veseli, M.Vranicar, R.Wellner, S.White, V.White

D0 Farm needs

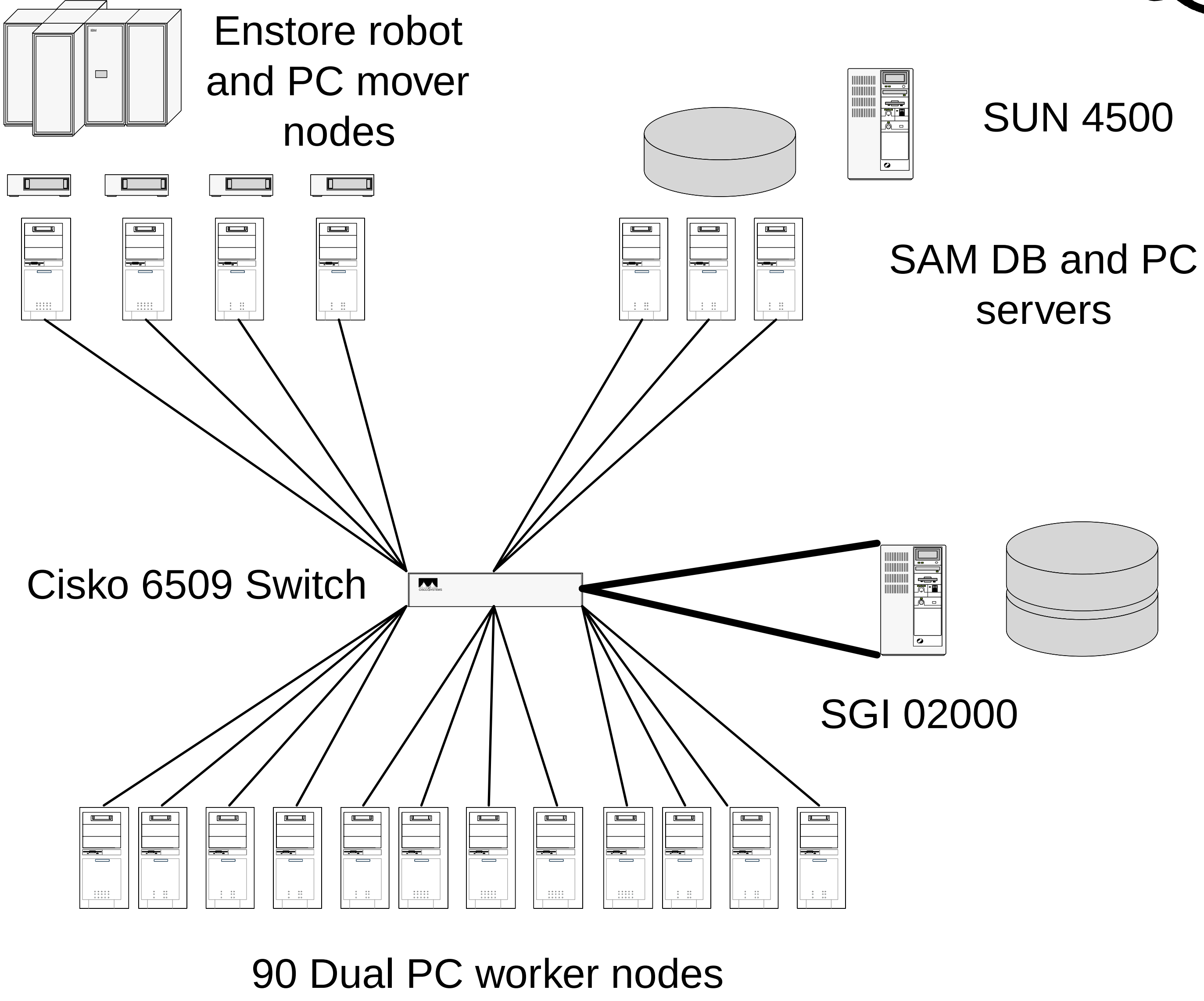
- 250K event size
- 50Hz trigger rate
  - peak rate of 12.5 MB/sec
  - DC is less but reprocessing will bring back up

Goal

Reality

- Reconstruction 5- 20 seconds/event on 750 MHz PIII
  - need 250->>500 CPU's to handle peak rate
  - DC is 40% of peak
  - time constant for 1 GB file is 5- 10 hours.

D0 Farms



Enstore robot and PC mover nodes

SUN 4500

SAM DB and PC servers

Cisko 6509 Switch

SGI 02000

90 Dual PC worker nodes

I/O machine

- Purpose
  - split/merge of farm output
  - Serve home areas
  - Batch system control
  - File delivery master
- D0bbin
  - 4 CPU SGI 02000
  - 2 GB ethernet cards
  - 4 72 GB disk partitions (2 way stripe)
  - peak I/O rates of 40-60 MB/sec

Worker Nodes

- 40 Dual Pentium III 500MHz
  - 256MB/CPU
- 50 Dual Pentium III 750MHz
  - 512 MB/CPU
- 2 data disks (18 GB) + 6GB system
- 100Mb ethernet
- CD/floppy for system configuration

Design Principles

- Use existing facilities
  - SAM/Enstore for data access and file tracking
  - Farm batch system (FBS) for most job control
- Keep D0 farm control scripts to a minimum
  - Batch system assigns machines
  - Data access system decides which file you get
- If worker process or machine dies, lose minimal number of files and don’t affect other processes
- No heroic recovery measures, track and resubmit those files

Worker Configuration

- Workers act as generic FNAL farm machines
  - Only customization is pnfs for file delivery, home area mount and startup of sam daemons on reboot.
  - D0 code environment downloads at job start
  - data access through SAM/encp/rcp, database server

- Batch system assigns workers to job, not D0FARM control process.
- D0FARM control never knows which workers are assigned to a job and does not need to.

Data Access is SAM/enstore

- Integrated data handling system
- File and process data base
- Data base server
- File servers
- Enstore File delivery systems
- Pnfs file system

Farm Perspective

Can tell it you want a set of files

Can ask for the 'next' file

Can flag file as processed or error

Can get detailed accounting on what happened

Data transfers are from ~ 12 mover nodes to 90 farm nodes through 6509 switch – theoretically could move 100's of MB/sec

Reality – online system has priority for drives.

Farm Batch System
Typical Farm Job

```
SECTION START
      EXEC=startjob  parameters
      QUEUE=D0bbin
SECTION WORKER
      EXEC=runjob  parameters
      NWORKERS=20
      QUEUE=D0worker
SECTION END
      EXEC=stopjob parameters
      QUEUE=D0bbin
      DEPEND WORKER(done)
```

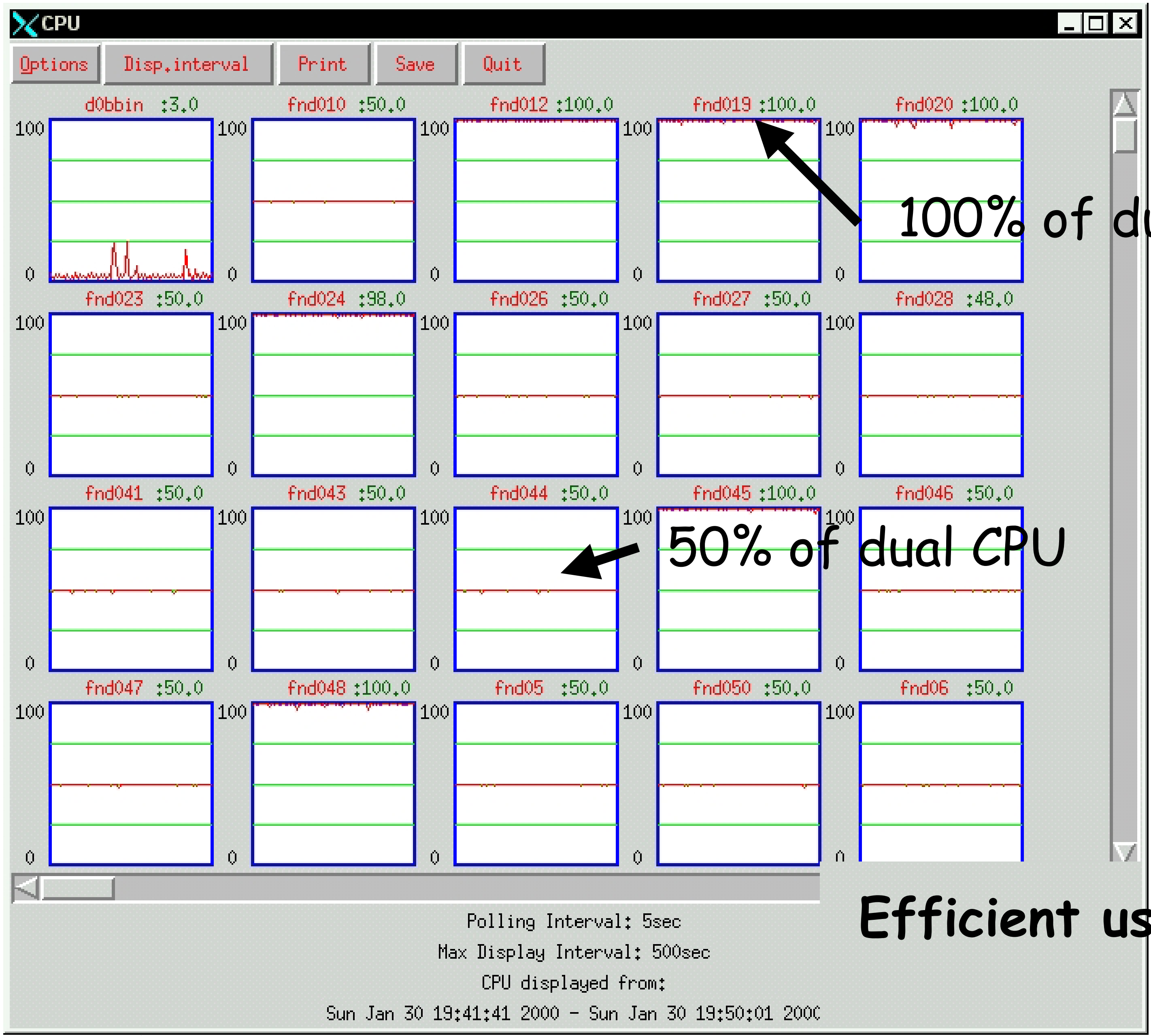
- Queue tells the system what kind of machine to run on and how many.
- EXEC gives the script name and parameters
- DEPEND allows cleanup section to run when all worker sections are done.
- FBS assigns temporary disk on workers
- On end yanks disk and kills all processes.

Structure of a Farm Job

- START:
  - Tell SAM which files you will want
  - Go into wait state until get end signal
- WORKER: runs on N nodes
  - Download D0 environment
  - Inform SAM ready for data
  - Ask for SAM for next file
  - Process file and store output to output buffer
  - Inform SAM of success and ask for next file
  - On error or end of list, terminate.
- END:
  - Create job summary
  - Send message to Start process telling it to shut down the SAM connection for input

Farm Batch System Monitor

D0 Farms

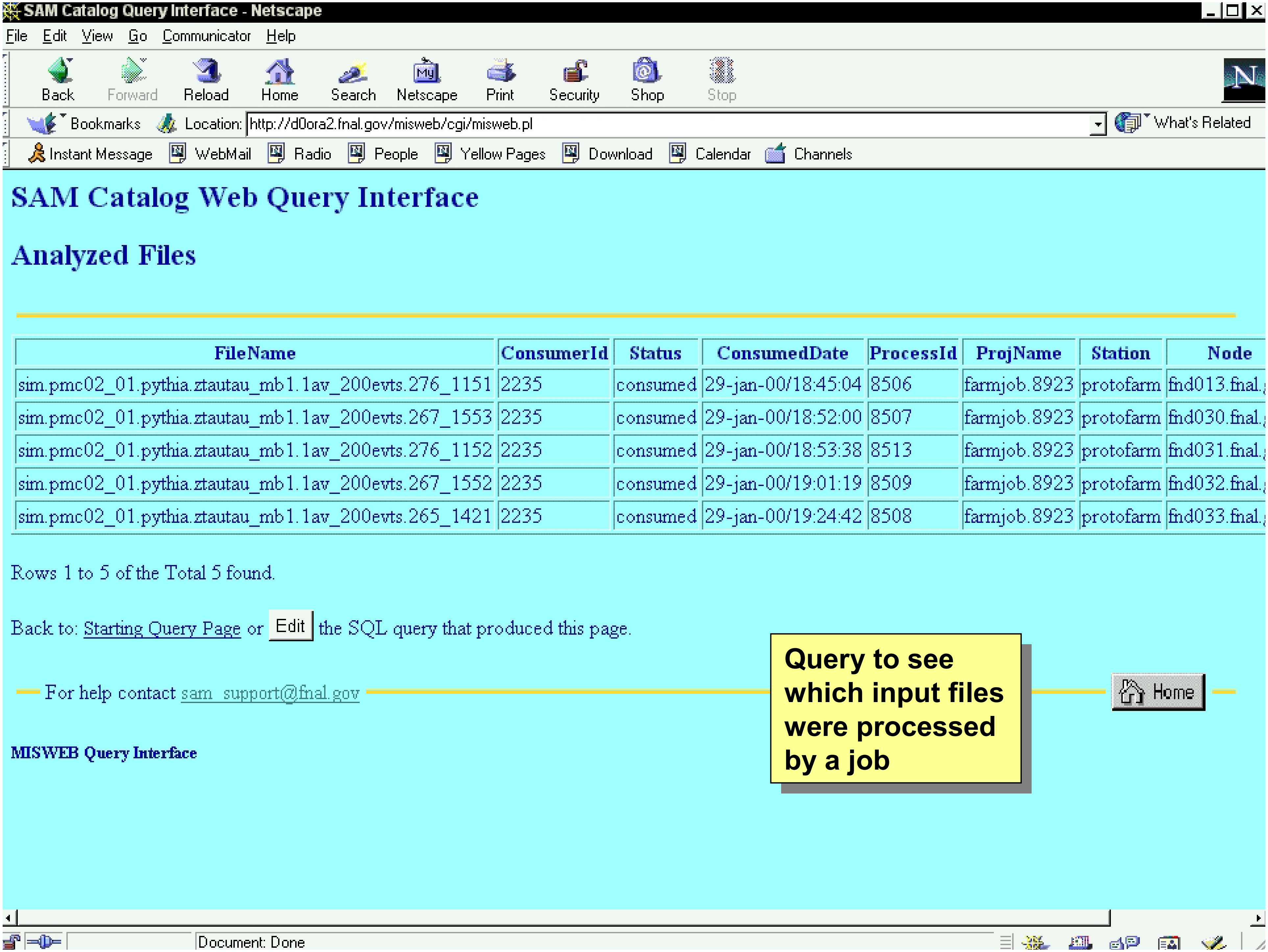


```
Polling Interval: 5sec
Max Display Interval: 500sec
CPU displayed from:
Sun Jan 30 19:41:41 2000 - Sun Jan 30 19:50:01 200C
```

100% of dual

50% of dual CPU

Efficient use of of CPU


SAM Catalog Query Interface - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://d0ora2.fnal.gov/misweb/cgi/misweb.pl
What's Related
Instant Message WebMail Radio People Yellow Pages Download Calendar Channels
Document: Done

SAM Catalog Web Query Interface

Analyzed Files

FileName	ConsumerId	Status	ConsumedDate	ProcessId	ProjName	Station	Node
sim.pmc02_01.pythia.ztautau_mb1.1av_200evts.276_1151	2235	consumed	29-jan-00/18:45:04	8506	farmjob.8923	protofarm	fnd013.fnal.
sim.pmc02_01.pythia.ztautau_mb1.1av_200evts.267_1553	2235	consumed	29-jan-00/18:52:00	8507	farmjob.8923	protofarm	fnd030.fnal.
sim.pmc02_01.pythia.ztautau_mb1.1av_200evts.276_1152	2235	consumed	29-jan-00/18:53:38	8513	farmjob.8923	protofarm	fnd031.fnal.
sim.pmc02_01.pythia.ztautau_mb1.1av_200evts.267_1552	2235	consumed	29-jan-00/19:01:19	8509	farmjob.8923	protofarm	fnd032.fnal.
sim.pmc02_01.pythia.ztautau_mb1.1av_200evts.265_1421	2235	consumed	29-jan-00/19:24:42	8508	farmjob.8923	protofarm	fnd033.fnal.

Rows 1 to 5 of the Total 5 found.

Back to: Starting Query Page or Edit the SQL query that produced this page.

For help contact sam_support@fnal.gov

Home

MISWEB Query Interface

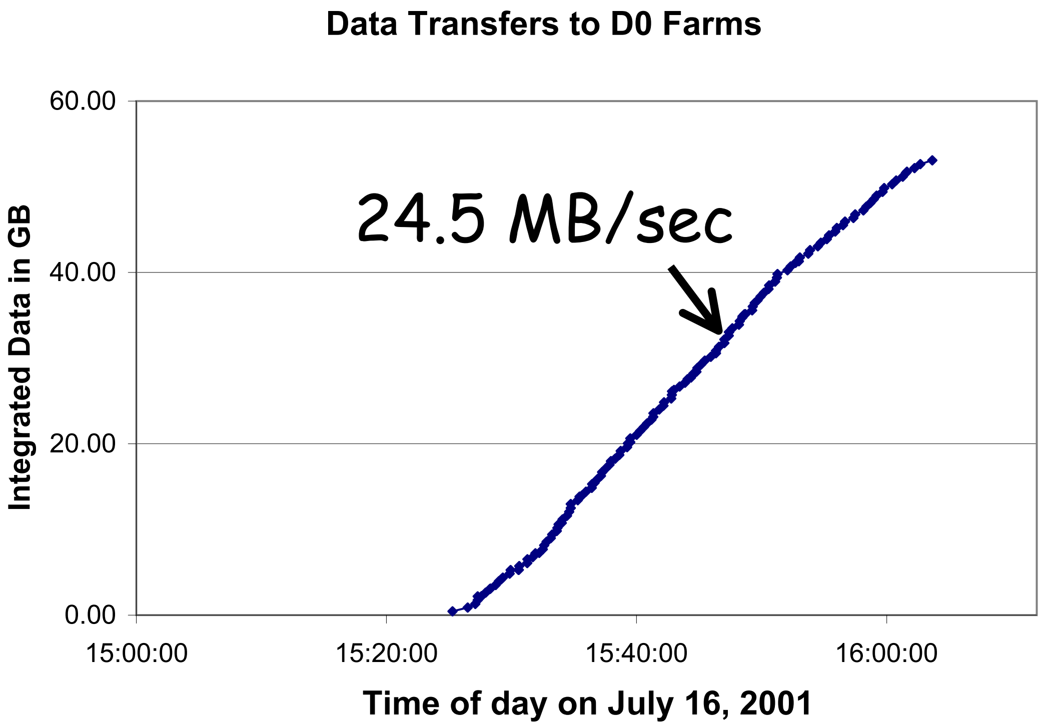
Query to see which input files were processed by a job

Status

- System has been in use for MC processing since before CHEP 2000
- System has been processing data as it comes off the D0 detector since March 2001
- Hardware/control/monitoring can handle full data rates well but…
- Major problem is speed of executable and data expansion during detector debugging
  - Output size is ~ input size by design
  - Currently factor of 2-3 larger due to debugging info.
  - Better thresholds and less noise will make life much easier

- *Farms get more stress at beginning of run than later!!*

Results of typical farm startup

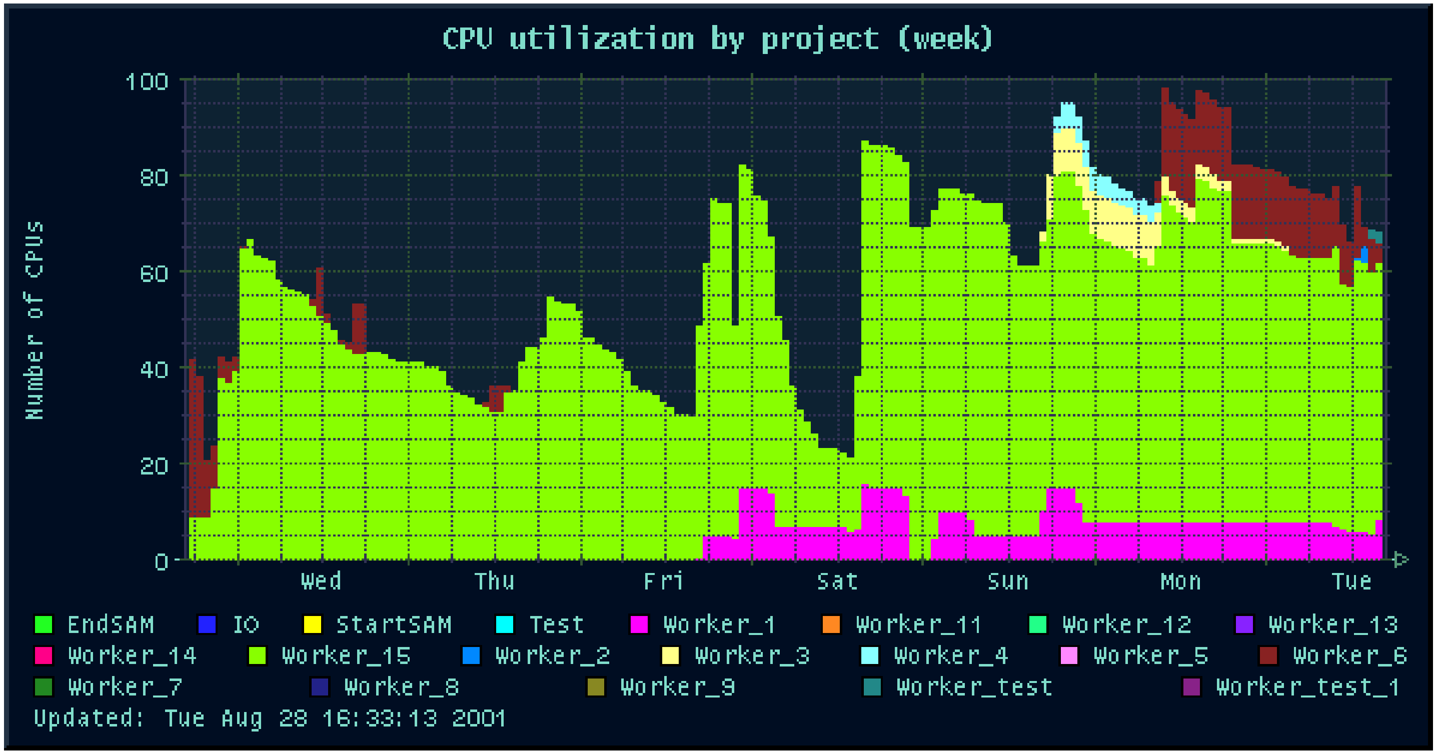


24.5 MB/sec

- Cold start of ½ of the D0 farm.
- 90 receiver nodes
- 141 files of average size 376 MB
- Read from 2-3 network mounted Mammoth II tapes over 100 MB ethernet at ~10MB/sec/drive.
- Elapsed time of 44 minutes.
- This is twice peak rate from the detector.

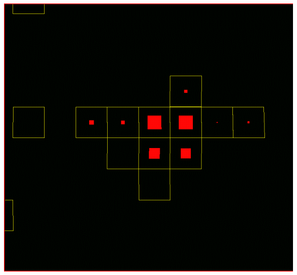
Current Production

http://www-isd.fnal.gov/cgi-bin/fbsng/fbswww/fbswww?action=graphs&period=week&farm=D0

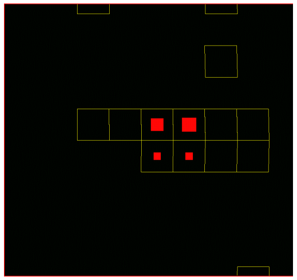


Plot from FBSWWW product – out of the box

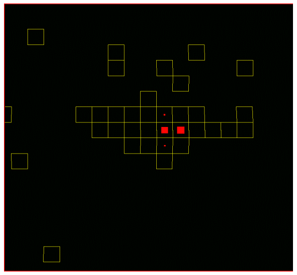
$W \rightarrow e\nu$ candidate



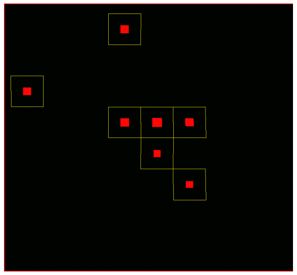
Layer 1
8 GeV



Layer 2
14 GeV



Layer 3
15 GeV

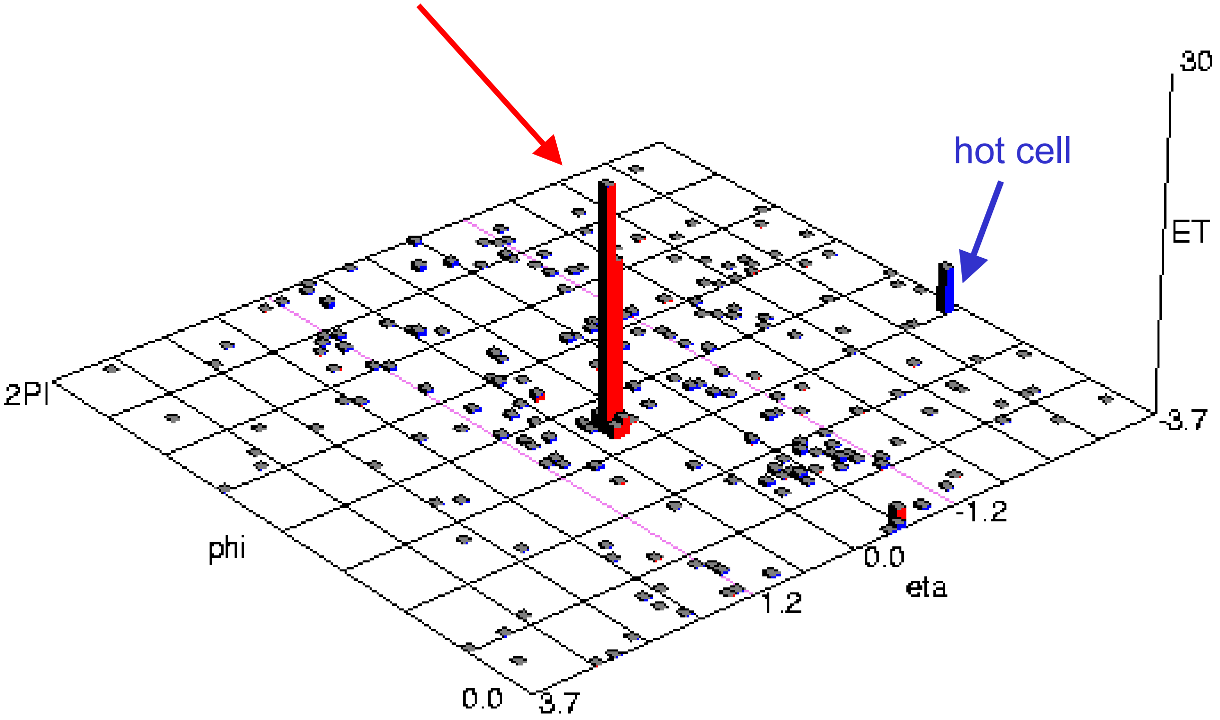


Layer 4
0.3 GeV

Run 125232 Event 183666

Electron candidate recorded using EM trigger

p_T = 38 GeV isolation<0.2 EM fraction = 0.97



Pierre Petroff and Laurent Duflot

Future

- It works now! but we will still:
- Add ~100 more nodes over next 6 months
- Make Improvements in automated running

    Datasets currently defined and submitted by hand

    ~ .25 FTE but still too much

- Local caching of files

    Guarantee tape streams at full speed

    Don't waste tape mounts if process file multiple times